Theoretical guidelines for a partially informed forecast examiner

Alexander Tsyplakov

Department of Economics, Novosibirsk State University

2. April 2014

Theoretical guidelines for a partially informed forecast examiner

Alexander Tsyplakov

Department of Economics, Novosibirsk State University

April 2, 2014

Abstract

The paper explores probability theory foundations behind evaluation of probabilistic forecasts. The emphasis is on a situation when the forecast examiner possesses only partially the information which was available and was used to produce a forecast. We argue that in such a situation forecasts should be judged by their conditional auto-calibration. Necessary and sufficient conditions of auto-calibration are discussed and expressed in the form of testable moment conditions. The paper also analyzes relationships between forecast calibration and forecast efficiency.

Key words: probabilistic forecast; forecast calibration; moment condition; probability integral transform; orthogonality condition; scoring rule; forecast encompassing.

JEL classification: C53; C52.

1 Introduction

There is little doubt that it is important for users of economic forecasts to have information on the degree of forecast uncertainty and probabilities of different scenarios. In general point forecasts give insufficient information to a user who needs to make a decision. This is the reason for growing popularity of more complete—probabilistic—forecasts in econometrics.

Real-life forecasts are not perfect and we want to be able to diagnose imperfections in order to improve our forecasting methods. Several procedures have been used for testing forecast calibration and efficiency in the literature on probabilistic forecasts. For example, see Kupiec (1995), Diebold, Gunther, and Tay (1998), Christoffersen (1998), Diebold, Tay, and Wallis (1999), Berkowitz (2001), Clements and Taylor (2003), Wallis (2003), Engle and Manganelli (2004), Clements (2006), Mitchell and Wallis (2011), Chen (2011), Galbraith and van Norden (2011) (this list includes closely related literature on interval/quantile forecasts). However, most of these procedures are applicable only in a narrow class of forecasting situations, primarily when one-step-ahead forecasts of a time series are made given the full previous history of this series. The literature does not provide a comprehensive general picture of testable implications of forecast calibration/efficiency. Even the conditions under which we can call a forecast calibrated or efficient are not yet fully understood and formally stated. In this paper we want to make up for this omission.

To catch the idea of the approach employed in this paper consider the example of the classical point forecasting under quadratic loss. It is well-known that the conditional expectation with respect to an information set Ψ is the best point forecast in mean-square sense of all

the Ψ-measurable forecasts. From the properties of conditional expectation it follows that the efficient forecast must be unbiased and the forecast error must be uncorrelated with any Ψ-measurable variables. These theoretical properties lead to corresponding test procedures, for example, Mincer-Zarnovitz-type regression-based tests (Mincer and Zarnowitz, 1969).

This paper applies a similar approach to probabilistic forecasts. Unlike most of the existing literature (Corradi and Swanson, 2006c is a vivid example) we emphasize the probability theory basis behind probabilistic forecasting, rather than statistical testing. It turns out that for many important theoretical results one can consider forecasting as a one-shot activity, rather than repeated one, which is subject to statistical procedures.

We consider forecasting of some target outcome Y which is a real-valued random variable. A complete probabilistic forecasts of Y is represented by a random cumulative distribution function $\tilde{F}$. (Tilde is used to show that the CDF is random). To analyze the properties of a probabilistic forecast it is necessary to consider the joint distribution of the forecast $\tilde{F}$ and the target outcome variable Y specified on a common probability space. Finally, forecast evaluation must rely on some relevant information represented by an information set Ψ. Formally Ψ is a sub-σ-algebra in the underlying probability space.

For a point forecast judged by the quadratic loss it is important to correctly represent the central point of the conditional distribution of Y given the relevant information set, which is achieved when the forecast coincides with the conditional mean. Similarly, for a probabilistic forecast it is important to be calibrated (Diebold, Hahn, and Tay, 1999; Gneiting, Balabdaoui, and Raftery, 2007). Calibration means good conformity between a probabilistic forecast and the actual behavior of the target variable. However, this idea of conformity is vague and requires an accurate formulation.

Several different modes of calibration were considered in the literature: probabilistic calibration (PIT uniformity), marginal calibration and ideal calibration with respect to an information set (Gneiting, Balabdaoui, and Raftery, 2007, Gneiting and Ranjan, 2013). Also very popular is the condition of uniformity and independence of PIT values (Diebold, Gunther, and Tay, 1998).

Many papers on evaluation of probabilistic forecasts assume that there is a complete parametric model of the data which gives the DGP for some true vector of parameter values (e.g. Corradi and Swanson, 2006a; Corradi and Swanson, 2006c; Chen, 2011). From this point of view forecast evaluation is a type of model evaluation. The probabilistic properties of forecast evaluation statistics are governed by the model. However, this approach is not applicable when there is no formal parametric model behind the forecasts. Our view is that the concept of calibration should not refer to a "model" or the "true DGP". It is more realistic to consider forecasting *methods* rather than forecasting *models* (cf. Giacomini and White, 2006). This permits us not to exclude forecasts which are not based on formal models (e.g. survey forecasts or forecasts utilizing exponential smoothing).

The mode of calibration closest to the idea of "true DGP" is ideal calibration (to be defined below). In a sense, it is fundamental, comprehensive and underlies the current econometric literature on probabilistic forecasting with its reliance on model-based forecasts. However, practical considerations suggest a different (though closely related) concept of calibration.

In a forecast evaluation situation one should distinguish (at least) two different parties: the forecaster and the individual who evaluates the forecast. The later party will be called the examiner here. The information sets of the forecaster and the examiner can be distinct, say, Ψ^* and Ψ. The concept of ideal calibration is ambiguous without specifying the information set. The forecast, which is calibrated with respect to Ψ^*, can be miscalibrated with respect to Ψ and vice versa.

There are many real-life situations in which Ψ^* and Ψ are not the same. For example, the central bank or the government can use internal information which is not publicly available. It is also not uncommon that a forecast includes subjective judgment of the forecaster or utilizes suboptimal sources of information. The leading motivating example is that of survey forecasts (like Survey of Professional Forecasters, see Diebold, Tay, and Wallis, 1999, Clements, 2006, Engelberg, Manski, and Williams, 2009). To be comprehensive enough the theory of forecast evaluation should not exclude the possibility that a forecaster uses some irrelevant information. (One can recall Roman augurs using observation of birds' behavior for foretelling.)

The idea is that it is in general more convenient to define a notion of calibration which explicitly takes into account the fact that the forecast itself can be the source of information for the forecast examiner. We call this mode of calibration *conditional auto-calibration*. The definition is given below. In the situation of point forecasting under quadratic loss the idea can be expressed as follows: a point forecast Y^f is efficient (or rational) if $\mathsf{E}[Y|\Psi, Y^f] = Y^f$.

The main reason for introducing the concept of conditional auto-calibration along with ideal calibration is that if forecaster's information set is not known to the examiner, the later has no way to verify that the forecast is ideally calibrated with respect to this information set. Hence the examiner in fact can only test auto-calibration.

For practical reasons it is convenient to express the implications of calibration in the form of moment conditions. Here we confront with a subtle point of probabilistic forecasting, namely, that there are two different kinds of moments involved. First, there are moments defined on the underlying probability space. Second, a realization of a complete probabilistic forecast $\tilde{F}$ specifies a probability measure for outcomes, which can be used to calculate various moments. If the forecast is calibrated and coincides with a relevant conditional distribution function, then the corresponding conditional moments can be expressed in terms of the moments calculated from the forecast-based probability measures. In this paper we show how different aspects of calibration can be characterized by the corresponding moment conditions. The key to many theoretical results mentioned in this paper is Theorem 6 in Appendix. It allows to express conditional moments given Ψ in terms of the corresponding moments of conditional distributions in the case when moment function includes Ψ-measurable random elements.

Another aspect of probabilistic forecasting is forecast efficiency (also called optimality and rationality in different contexts) with respect to the objectives of a forecast user. In a sense calibration is a testable implication of efficiency. However we also consider more direct conditions of efficiency and relate them to calibration. Among other things, we formulate and prove the sharpness principle of forecasting conjectured in Gneiting, Balabdaoui, and Raftery (2007). Its formulation happens to rely on the notion of auto-calibration.

The question of accuracy testing, which also can be used for forecast evaluation, is not considered here. There is an extensive literature on this subject starting from Diebold and Mariano (1995) and including West (1996), White (2000), Sarno and Valente (2004); Corradi and Swanson (2005), Corradi and Swanson (2006b), Giacomini and White (2006), Amisano and Giacomini (2007), Bao, Lee, and Saltoğlu (2007). However, we state relative predicted efficiency conditions which are closely related and can be viewed in some sense as an improved alternative to accuracy testing when it is used for the purpose of calibration or efficiency testing.

Section 2 analyzes the notion of calibration and characterize calibration by moment conditions. Section 3 discusses forecast efficiency from the point of view of utility maximization and proper scoring rules and analyzes the links between calibration and efficiency. Section 4 provides illustrative examples. Section 5 compiles main conclusions. Theorems and a technical counterexample are moved to the Appendix.

2 Forecast calibration

2.1 Basic definitions

We start by defining ideal calibration.[1] Assume that Ψ is the information set including all the relevant information which can be used. The idea is that for a given information set Ψ the ideally calibrated forecast, first, fully utilizes Ψ and, second, is based only on Ψ without employing any other information (formally, the forecast $\tilde{F}$ is Ψ-measurable). The arguments in Subsection 3.1 below give additional reason for using the word "ideal" in this context: in a sense it is the best achievable forecast among forecasts based on Ψ.

Definition 1. A forecast $\tilde{F}$ is *ideally calibrated given* Ψ if $\tilde{F} = \mathbb{F}_\Psi$, where $\mathbb{F}_\Psi(y) = \mathbb{F}(y|\Psi)$ is the distribution function of Y conditional on Ψ.

If the forecaster possesses some information which is not available to the examiner, then the examiner can potentially derive some new information from the forecast itself. Thus, in this case the relevant information set to be used for forecast evaluation combines the examiner's prior information with the information delivered by the forecast and we can state that forecast $\tilde{F}$ is calibrated from the examiner's point of view if it coincides with $\mathbb{F}_{\Psi,\tilde{F}}$, which is the conditional distribution of Y given Ψ and $\tilde{F}$.

Definition 2. A forecast $\tilde{F}$ is *conditionally auto-calibrated given* Ψ if it is ideally calibrated with respect to $\Psi \cup \sigma(\tilde{F})$, that is, $\tilde{F} = \mathbb{F}_{\Psi,\tilde{F}}$.

The definition extends the notion of (unconditional) auto-calibration introduced in Tsyplakov (2011).

By definition all auto-calibrated forecasts are ideally calibrated given the corresponding information set $\Psi \cup \sigma(\tilde{F})$. Moreover, a Ψ-measurable forecast, which is auto-calibrated given Ψ, must be ideally calibrated given Ψ. Conversely, it can be stated that any forecast, which is ideally calibrated with respect to some information set Ψ^* including Ψ, is auto-calibrated with respect to Ψ (Theorem 11). Therefore, if $\tilde{F}$ is auto-calibrated given Ψ_1 and Ψ_1 is a "richer" information set than Ψ_2, that is, Ψ_1 contains all the information of Ψ_2 and maybe some additional useful information (formally, $\Psi_2 \subset \Psi_1$), then it is auto-calibrated given Ψ_2.

Of course, one can base the theory of forecast evaluation on the definition of ideal forecast calibration. However, it is more clear and natural to concentrate on the property of conditional auto-calibration instead as this allows to highlight the specific aspects of a situation when the forecast examiner can obtain from the evaluated forecast itself information which is new to him.

As discussed below, in general it is not sufficient to use conditions of PIT uniformity and orthogonality between PIT values and some observable variables based on Ψ to test calibration (which can be put into regression to test orthogonality conditions). Thus, a partially informed examiner confronted with a black-box forecast have to use specific instruments and construct peculiar variables based on both $\tilde{F}$ and Ψ, which can be used in calibration testing.

Even if the forecast under examination is known to be Ψ-measurable, these specific instruments can be utilized with benefit, because for the examiner it might not be clear how the forecast is constructed from Ψ. Moreover, even if the forecast is not a black box one these specific

[1] The concept of ideal calibration with respect to an information set is quite natural and is implicit in the literature on probabilistic forecasting (albeit, possibly, in a non-direct fashion—like "forecast represents the true DGP"). Cf. Diebold, Gunther, and Tay (1998). Explicit definitions can be found in Tsyplakov (2011) and Gneiting and Ranjan (2013). It is also similar to the definition of interval forecast efficiency with respect to an information set in Christoffersen (1998).

instruments can be useful, because at the technical side a forecast $\tilde{F}$ is not a finite-dimensional variable which is a typical object of analysis in econometrics. There are specific aspects of using a CDF-valued variables, and we illustrate these in examples below.

A final remark is pertinent here. It goes without saying that adequate choice of information set is crucial for testing calibration in applications. If an examiner wants to evaluate a forecast he must consider information Ψ which is available at the time the forecast was made. Further, judgments about forecaster's rationality can only be based on the information available to this forecaster.

2.2 General moment conditions of forecast calibration

The definition of conditional auto-calibration with respect to an information set, although intuitively appealing, is too abstract. For the purposes of forecast evaluation one would like to have some functions of Y, which are directly observable and could be compared to something, which is based on the forecast and the information contained in Ψ. In the case of point forecasting we can directly compare the forecasts and the actual realizations of Y. We would like to have something similar in the case of probabilistic forecasting.

Consider a function $g(y, w, F)$, which takes an outcome value y, a distribution function F and some additional variable w as its arguments. A good probabilistic forecast of Y should be able to predict the behavior of $g(Y, W, \tilde{F})$, where W is some Ψ-measurable random element. Here $\tilde{F}$ and W can be treated as fixed, since they are assumed to be already known at the forecasting time. Thus, letting $\tilde{F} = F$ and $W = w$ we can use the expectation of $g(Y, w, F)$ under the assumption that Y is distributed according to F as our forecast of $g(Y, W, \tilde{F})$.

Note that we have to consider two types of probability measures and two types of expectations here. First, there is a probability measure in the underlying probability space. Second, values, which are assumed by forecasts, are CDFs inducing their own probability measures.

This reasoning suggests a very general type of moment condition of calibration. Define

$$\gamma(F, w) = \mathsf{E}_{Y \sim F} g(Y, w, F).$$

Then we must have under conditional auto-calibration that

$$\mathsf{E} g(Y, W, \tilde{F}) = \mathsf{E}\gamma(\tilde{F}, W). \tag{1}$$

Actually this is the most general moment condition of calibration, because, as discussed below, it is not only necessary, but also sufficient for conditional auto-calibration (see subsection 2.7). The conditions of this kind can be used to compare theoretical and empirical moments for the purpose of calibration testing (see subsection 2.8).

As conditions (1) are rather general, it is interesting to narrow these testable implications of calibration and consider various special cases. We start by linking it to the conditions of probabilistic and marginal calibration.

2.3 Conditional probabilistic and marginal calibration

Probabilistic calibration (PIT uniformity) Consider a situation when forecasts are such that their values have a constant support $[a, b]$ with possibly infinite bounds, continuous and strictly increasing at $[a, b]$. (The outcome variable Y is implicitly assumed to have the CDF with similar properties). We loosely call this setting a density forecasting situation. For a random variable Y with a cumulative distribution function F the probability integral transform (PIT)

value is defined as $F(Y)$. It has the $U[0,1]$ distribution if F is continuous. In the same manner one can define the PIT value for a probabilistic forecast.[2]

In a density forecasting situation the PIT value for a forecast $\tilde{F}$ and the outcome variable Y is defined as

$$P = \tilde{F}(Y).$$

This quantity is the one that is used most often for calibration diagnostics in econometrics; e.g. Diebold, Gunther, and Tay (1998); Mitchell and Wallis (2011); Chen (2011). Probabilistic calibration is a mode of calibration based on these PIT values. The term "probabilistic calibration" for the condition of PIT uniformity was suggested in Gneiting, Balabdaoui, and Raftery (2007); see also the reformulation of this definition in Gneiting and Ranjan (2013).[3] Here we introduce conditional version of this mode of calibration.

Definition 3. A forecast $\tilde{F}$ is *probabilistically calibrated given* Ψ if $P|\Psi \sim U[0,1]$.

This condition can be decomposed into two conditions, namely, that, first, PIT values are unconditionally distributed as $U[0,1]$ and, second, P and Ψ are independent (see Theorem 9).

Unconditional PIT uniformity can be assessed, for example, with the help of a histogram of the PIT values on the $[0,1]$ interval. The histogram should be almost flat (e.g. Diebold, Gunther, and Tay, 1998).

It can be seen that the concept of probabilistic calibration is closely connected to interval forecasting and quantile forecasting (e.g. value-at-risk forecasting).[4] For a forecast $\tilde{F}$ we can define the corresponding p-quantile forecast $Q_p = \tilde{F}^{-1}(p)$. Under correct calibration probability that Y is less than Q_p is p. Consequently the probability that P does not exceed $\tilde{F}(Q_p) = p$ should also be equal to p.

Further references and examples of moment conditions of probabilistic calibration can be found in Chen (2011). Under probabilistic calibration given Ψ all such moment conditions must be true also conditionally on Ψ.

More generally, for a function $c(p,w)$ taking $p \in [0,1]$ and an additional variable w as its arguments denote

$$\chi(w) = \mathsf{E}_{P\sim U[0,1]}\, c(P,w).$$

It's a direct corollary to Theorem 6 that if a forecast $\tilde{F}$ in a density forecasting situation is conditionally probabilistically calibrated with respect to Ψ (that is, $\tilde{F}(Y)|\Psi \sim U[0,1]$) then for any c any Ψ-measurable W it satisfies

$$\mathsf{E}c(\tilde{F}(Y),W) = \mathsf{E}\chi(W). \tag{2}$$

Note that these moment conditions are weaker than the general moment conditions of conditional auto-calibration (1).

[2]The notion of PIT can be extended to arbitrary distributions by introducing randomization (e.g. Ferguson, 1967; Brockwell, 2007). One can extend the results of the current paper in this direction, but we prefer not to do so in order to keep the exposition more transparent.

[3]The definitions of probabilistic and marginal calibration proposed in Gneiting, Balabdaoui, and Raftery (2007) are formulated from a prequential perspective due to Dawid (1984) for sequences of forecasts. In Gneiting and Ranjan (2013) the one-shot view on the forecasting theory is employed, similar to that of the current paper.

[4] That is why the literature in this area such as Kupiec (1995), Christoffersen (1998), Lopez (1998), Clements and Taylor (2003), Engle and Manganelli (2004) can be considered as a part of the literature on probabilistic forecasting.

Marginal calibration The concept of probabilistic calibration implicitly assumes a situation when probabilities are fixed while the bounds are reported by the forecaster. A reversed situation is when bounds are fixed while the forecaster reports probabilities as in the Survey of Professional Forecasters. A calibrated forecast must supply probabilities which are in accordance with the true ones. Probabilities for all possible bounds are summarized by a CDF. Thus, another mode of calibration is defined in terms of CDFs. The definition is given in Gneiting, Balabdaoui, and Raftery (2007) and reformulated in Gneiting and Ranjan (2013). Again, here we define a conditional version of this definition. In this definition E_Ψ is the operator of conditional expectation given Ψ and $\mathbb{F}_\Psi$ is the conditional CDF of Y given Ψ.

Definition 4. A forecast $\tilde{F}$ is marginally calibrated given Ψ if $\mathsf{E}_\Psi \tilde{F} = \mathbb{F}_\Psi$.

Similarly to conditional probabilistic calibration conditional marginal calibration can be characterized by moment conditions, which are weaker than the general moments conditions of conditional auto-calibration (1). For a function $n(y, w)$ of an outcome value y and an additional variable w denote

$$\nu(F, w) = \mathsf{E}_{Y\sim F} n(Y, w).$$

If a forecast $\tilde{F}$ is marginally calibrated with respect to Ψ then for any n any Ψ-measurable W it satisfies

$$\mathsf{E} n(Y, W) = \mathsf{E}\nu(\tilde{F}, W) \tag{3}$$

(see Theorem 10). That is, conditional marginal calibration implies that $\nu(\tilde{F}, W)$ is an unbiased forecast of $n(Y, W)$.

For example, one can take $n = y$, $\nu = \text{mean}(\tilde{F})$ to express mean unbiasedness of $\tilde{F}$. Gneiting, Balabdaoui, and Raftery (2007) propose a diagnostic diagram for unconditional marginal calibration based on binning and application of condition $\mathsf{E}\mathrm{I}\{Y \in (a, b]\} = \mathsf{E}[\tilde{F}(b) - \tilde{F}(a)]$ to the bins.

Theorem 8 states that both probabilistic and marginal calibration given Ψ are implied by auto-calibration given Ψ.[5] However, as we show below (subsection 2.7), neither probabilistic, nor marginal calibration are sufficient for auto-calibration with respect to the same information set. Hence, calibration tests based on (2) and (3) can be incomplete as tests of conditional auto-calibration.

2.4 Orthogonality conditions of calibration

From the theory of point forecasting it is known that the expectation conditional on the information set Ψ is the forecast which is optimal in mean-square sense among the forecast based on Ψ (e.g. Bierens, 2004, pp. 80–81). This forecast satisfies orthogonality conditions: the prediction error is uncorrelated with any random variable based on Ψ.[6] There are also extensions to the case of general cost functions (e.g. Granger, 1999). In Mitchell and Wallis (2011) an idea was put forward that calibration of probabilistic forecasts can be tested by verifying similar orthogonality conditions. We demonstrate that this idea lends itself to further generalization.

[5] Gneiting and Ranjan (2013) observe that the ideally calibrated forecast is both (unconditionally) marginally calibrated and probabilistically calibrated.

[6] These conditions were utilized in the rational expectations literature. Shiller (1978), p. 7: "…Expected forecast errors conditional on any subset of the information available when the forecast was made, are zero… Hence, the forecast error … is uncorrelated with any element of I_t [the set of public information available at time t]".

The term "orthogonality conditions" is known from the GMM literature (cf. Hansen, 1982).

Consider a function $r(y, F)$, which takes an outcome value y and a CDF F as its arguments, and denote

$$\rho(F) = \mathsf{E}_{Y\sim F} r(Y, F).$$

Let $a(w, F)$ be some function of an additional variable w and a CDF F.[7] By letting $g = ra$ in the general moment conditions of calibration (1) we obtain the following general orthogonality conditions of calibration: if a forecast $\tilde{F}$ is conditionally auto-calibrated with respect to Ψ then for any r, any Ψ-measurable W and any a it satisfies

$$\mathsf{E}[(r(Y, \tilde{F}) - \rho(\tilde{F}))\, a(W, \tilde{F})] = 0. \tag{4}$$

According to this conditions a point forecast ρ derived from a probabilistic forecast $\tilde{F}$ must be unbiased as a forecast of r and the error must not be correlated with any function of a Ψ-measurable W and the forecast $\tilde{F}$. One can represent a CDF F in function $a(w, F)$ by some characteristics of the corresponding distribution such as the mean, median or interquartile range.

An example of this type of orthogonality conditions can be found in Clements (2006), where in the context of evaluating the SPF probabilistic forecasts it was noted that $\mathsf{E}[(I - p)\, p] = 0$, where I is an indicator variable for the event that Y is in some interval and p is the predicted probability of this event.

Note that conditional probabilistic and marginal calibration can also be characterized by orthogonality conditions, but these conditions are less general. Under conditional probabilistic calibration with respect to Ψ for any function $k(p)$ taking $p \in [0, 1]$ as its argument and any Ψ-measurable W we must have

$$\mathsf{E}[(k(\tilde{F}(Y)) - \kappa) W] = 0, \tag{5}$$

where $\kappa = \mathsf{E}_{P\sim U[0,1]} k(P)$. Similarly under conditional marginal calibration with respect to Ψ for any function $m(y)$ of an outcome value y and any Ψ-measurable W we must have

$$\mathsf{E}[(m(Y) - \mu(\tilde{F}))\, W] = 0, \tag{6}$$

where $\mu(F) = \mathsf{E}_{Y\sim F} m(Y)$.

As an example of orthogonality conditions for probabilistic calibration consider a regression from subsection 4.3 of Christoffersen (1998) used for testing for conditional coverage of an interval forecast. A similar regression representing orthogonality conditions for marginal calibration can be found in Clements (2006).

2.5 Sequential auto-calibration

Consider a sequence $\tilde{F}_t$ of h-step-ahead probabilistic forecasts of a univariate time series Y_t, $t = 1, 2, \ldots$ in a recursive setting. The examiner's information set available for evaluating $\tilde{F}_t$, which we denote Ψ_t, should include information on available previous values of the series and previously issued forecasts. We assume that all forecasts $\tilde{F}_1, \ldots, \tilde{F}_t$ are already known at time $t - h$ and thus

$$\sigma(Y_1, , \ldots, Y_{t-h}, \tilde{F}_1 \ldots, , \tilde{F}_{t-1}) \subset \Psi_t.$$

This suggests the following definition.

[7] Note that any random element A which is measurable with respect to $\Psi \cup \sigma(\tilde{F})$ can be represented as $A = a(W, \tilde{F})$ for some function $a(w, F)$, where W is a Ψ-measurable variable.

Definition 5. A sequence of forecasts $\tilde{F}_t,\ t = 1,\dots,T$ in recursive h-step density forecasting situation is *sequentially auto-calibrated* if each forecast $\tilde{F}_t$ is conditionally auto-calibrated with respect to $\sigma(Y_1,,\dots,Y_{t-h},\tilde{F}_1\dots,,\tilde{F}_{t-1})$.

If a sequence of one-step density forecast of a time series $Y_t,\ t = 1,\dots.,T$ is made from the full history of the same series, then calibration is frequently judged by analyzing the resulting series of PIT values

$$P_t = \tilde{F}_t(Y_t), \qquad t = 1,\dots.,T.$$

It is assumed that a sequence of such forecasts is calibrated if and only if the PIT values P_t are independent and distributed as $U[0,1]$ (cf. Diebold, Gunther, and Tay, 1998). We will call this the UIPIT condition (condition of uniformity and independence of PIT values):

$$(P_1,\dots,P_T) \sim U[0,1]^T. \tag{UIPIT}$$

The UIPIT condition is very popular in the density forecast evaluation literature; e.g. Dawid (1984), Diebold, Gunther, and Tay (1998), Berkowitz (2001), Mitchell and Wallis (2011), Chen (2011). Mitchell and Wallis (2011) even call this "complete calibration". However, this is in fact not an independent mode of calibration. It can be a necessary condition of sequential auto-calibration or a sufficient condition of sequential ideal calibration (see subsection 2.7 below) or completely irrelevant depending on the situation.

UIPIT condition should be primarily considered as a necessary condition of sequential auto-calibration in a specific setting. If in a recursive one-step-ahead density forecasting situation forecasts $\tilde{F}_t,\ t = 1,\dots,T$ are sequentially auto-calibrated, then according to Theorem 12 UIPIT condition must hold. This is a generalization of Proposition in Diebold, Gunther, and Tay (1998), p. 867 in the spirit of partial information approach of the current paper.

The condition of serial independence of PIT values, which is the part of the UIPIT condition, can be expressed with the help of orthogonality conditions. For example, any function k of a the PIT value for moment t must be uncorrelated with any function k_2 of lagged PIT values:

$$\mathsf{E}[(k(P_t) - \kappa)k_2(P_{t-s})] = 0, \qquad s = 1,2,\dots$$

Under the UIPIT condition a series of transformed PIT values $k(P_t),\ t = 1,\dots,T$ also must be i.i.d. and hence serially uncorrelated. Therefore, we can use autocorrelation functions of the PIT values and their transformations to test sequential auto-calibration of recursive forecasts as proposed in Diebold, Gunther, and Tay (1998). When the UIPIT condition is not applicable, we can still employ similar orthogonality conditions

$$\mathsf{E}[(k(P_t) - \kappa)W_t] = 0$$

provided that we use only Ψ_t-measurable variables as W_t. For example, in the case of h-step-ahead forecasting we can use $W_t = k_2(P_{t-s})$ for $s \geq h$. In the case of real-time forecasting if some preliminary estimates of Y_{t-s}, say Y^*_{t-s}, are observed at the time when the forecast is made, then we can use

$$\mathsf{E}[(k(\tilde{F}_t(Y_t)) - \kappa)k_2(\tilde{F}_{t-s}(Y^*_{t-s}))] = 0, \qquad s = 1,2,\dots$$

In general there is no need to rely only on orthogonality conditions based on PIT values. Other conditions can be more suitable in many forecasting situations. It can be emphasized in particular that various functions of forecasts $\tilde{F}_1,\dots,\tilde{F}_t$ can be important in testing sequential auto-calibration.

2.6 Forecast encompassing

Next we consider forecast encompassing as an example of conditions of general type (1). The idea is to verify calibration of one forecasting method against another one.

Suppose that we want to test whether $\tilde{F}_1$ is auto-calibrated and $\tilde{F}_2$ is an alternative forecast. Forecast examiner can use an information set Ψ and information contained in forecast $\tilde{F}_2$ for forecast evaluation purposes. For a pair of (non-random) CDFs H and F and some additional variable w let

$$\gamma_0(H, w, F) = \mathsf{E}_{Y\sim H} g(Y, w, F).$$

Then for two forecasts $\tilde{F}_1$, $\tilde{F}_2$ under the assumption that $\tilde{F}_1$ is auto-calibrated with respect to $\Psi \cup \sigma(\tilde{F}_2)$ we have for any g and any Ψ-measurable W

$$\mathsf{E} g(Y, W, \tilde{F}_2) = \mathsf{E}\gamma_0(\tilde{F}_1, W, \tilde{F}_2).$$

This can be called a *forecast encompassing* condition. The idea applying encompassing principle to forecasts is due to Chong and Hendry (1986). The principle states that "models which claim to congruently represent a data generation process must be able to account for the findings of rival models" (Chong and Hendry, 1986, p. 676).

We can note here that full probabilistic forecasts are particularly suited for application of the encompassing principle since they provide *complete* distribution functions, so that given one probabilistic forecast we can derive forecast of *any* calibration-related characteristics of another probabilistic forecast.

In particular, for $g = k(F(y))a(W, F)$ we obtain

$$\mathsf{E}[(k(\tilde{F}_2(Y)) - \kappa_0(\tilde{F}_1, \tilde{F}_2))a(W, \tilde{F}_2)],$$

where $\kappa_0(H, F) = \mathsf{E}_{Y\sim H} k(F(Y))$. As an example of this forecast encompassing condition consider an indicator function $k = \mathrm{I}\{F(Y) \le p\} = \mathrm{I}\{Y \le F^{-1}(p)\}$. The expectation for $F = F_2$ under another CDF F_1 is given by $\kappa_0(F_1, F_2) = F_1(F_2^{-1}(p))$ and we obtain the following moment condition:

$$\mathsf{E}[(\mathrm{I}\{\tilde{F}_2(Y) \le p\} - \tilde{F}_1(\tilde{F}_2^{-1}(p)))a(W, \tilde{F}_2)] = 0.$$

Similarly for $g = m(y)a(W, F)$ we obtain an orthogonality condition

$$\mathsf{E}[(m(Y) - \mu(\tilde{F}_1))a(W, \tilde{F}_2)].$$

Another form of forecast encompassing contrasts results of one forecast with results of another one. The idea is that a calibrated forecast $\tilde{F}_1$ must be able to explain the differential in some function r for forecasts $\tilde{F}_1$ and $\tilde{F}_2$. When $\tilde{F}_1$ is well-calibrated we have

$$\mathsf{E}[g(Y, W, \tilde{F}_2) - g(Y, W, \tilde{F}_1)] = \mathsf{E}[\gamma_0(\tilde{F}_1, W, \tilde{F}_2) - \gamma(W, \tilde{F}_1)],$$

where $\gamma(w, F) = \gamma_0(F, w, F)$. The two forms of forecast encompassing conditions roughly correspond to FE(2) and FE(3) regressions in Clements and Harvey (2010) where forecast encompassing is applied to probability forecasts of 0/1 events.

2.7 Sufficient conditions of calibration

In some sense general moment conditions (1) are complete. That is, it can be proved (see below) that the requirement that they are satisfied for any g and any W is sufficient for conditional

auto-calibration. In the same sense conditions (2) and (3) are sufficient for conditional probabilistic and marginal calibration respectively. However, it is tempting to narrow these general moment conditions somehow. Both theoretically and practically interesting question is how "narrow" one can be in calibration testing without a fundamental sacrifice of comprehensiveness.

We have already seen that auto-calibration given Ψ implies both probabilistic and marginal conditional calibration given Ψ. Probabilistic calibration and marginal calibration are different concepts. Neither of them generalizes the other one. Counterexamples[8] for a density forecasting situation and trivial Ψ can be found in Gneiting, Balabdaoui, and Raftery (2007) (Examples 3, 5, 6) and in Mitchell and Wallis (2011) (combined and unfocused forecasts in the AR(2) example). See also Forecast C in Example 1 below.[9]

It can be seen that neither probabilistic, nor marginal calibration given Ψ are sufficient for auto-calibration given Ψ. Example 13 in the Appendix demonstrates that even when a forecast is simultaneously (unconditionally) probabilistically and marginally calibrated, it can fail to be auto-calibrated.

Of course, if we do not want to assume that the forecast examiner is only partially informed, then the distinction above is not important. If a Ψ-measurable forecast $\tilde{F}$ is conditionally marginally calibrated with respect to Ψ (that is, $\mathsf{E}_\Psi \tilde{F} = \mathbb{F}_\Psi$) then obviously $\tilde{F}$ is ideally calibrated with respect to Ψ. The same is true for conditional probabilistic calibration (Theorem 14).

Both probabilistic and marginal calibration with respect to $\Psi \cup \sigma(\Psi)$ are equivalent to auto-calibration given Ψ. So are orthogonality conditions (7) and (8) which follow. They are arguably the most narrow sufficient conditions of conditional auto-calibration.

Marginal calibration with respect to $\Psi \cup \sigma(\tilde{F})$ can be expressed in terms of orthogonality conditions between $\mathrm{I}\{Y \le y\} - \tilde{F}(y)$ and any function $a(W, \tilde{F})$ of a Ψ-measurable W and forecast $\tilde{F}$ (for any real y and any a). Similarly (in a density forecasting situation) probabilistic calibration with respect to $\Psi \cup \sigma(\tilde{F})$ can be expressed in terms of orthogonality conditions between $\mathrm{I}\{\tilde{F}(Y) \le p\} - p$ and any $a(W, \tilde{F})$. Thus, we have two different sufficient moment conditions of auto-calibration with respect to Ψ:

$$\mathsf{E}[(\mathrm{I}\{Y \le y\} - \tilde{F}(y))\, a(W, \tilde{F})] = 0 \tag{7}$$

for any real y, any a and any Ψ-measurable W and

$$\mathsf{E}[(\mathrm{I}\{\tilde{F}(Y) \le p\} - p)\, a(W, \tilde{F})] = 0 \tag{8}$$

for any $p \in [0, 1]$, any a and any Ψ-measurable W.

There is no guarantee that these partial conditions with indicator variables can provide tests with good power. Perhaps, some more general test based on conditions (1) can be more powerful. At least an examiner utilizing such narrow sufficient conditions would not be fundamentally non-comprehensive.

Theorem 15 states a less obvious sufficient moment condition of conditional auto-calibration:

$$\mathsf{E}[(r(Y, \tilde{F}) - \rho(\tilde{F}))\, W] = 0, \tag{9}$$

[8] One can easily generate other counterexamples. In theory an arbitrary forecast can be readily recalibrated to achieve either probabilistic or marginal calibration relative to Ψ. If $\tilde{G}(p)$ is the conditional distribution function of PIT values $\tilde{F}(Y)$ given Ψ, then $\tilde{G}(\tilde{F}(y))$ is a probabilistically recalibrated version of $\tilde{F}(Y)$. Similarly $\tilde{F}(\tilde{H}^{-1}(\mathbb{F}_\Psi(y)))$, where $\tilde{H}(y) = \mathsf{E}_\Psi \tilde{F}(y)$, is its marginally recalibrated version.

[9] Probabilistic and marginal calibration are also distinct concepts for discrete Y assuming more than two values; see an example in Table 2 of Gneiting and Ranjan (2013).

for any r and any Ψ-measurable W. This is also an orthogonality condition, where orthogonality is between $r(Y,\tilde{F})-\rho(\tilde{F})$ and Ψ-measurable variables.

One can never be sure that a forecast is conditionally auto-calibrated (probabilistically calibrated, marginally calibrated). All of the theorems on sufficient moment conditions require corresponding conditions to be satisfied for arbitrary functions and arbitrary variables W. Comparison of forecasts A and B in Example 1 below highlights this problem.

Our view is that the problem is a fundamental one and there is no universal solution. However, we can give a universal advise for a forecast examiner: try to build as good forecast as you yourself can or find some other good forecast and use relative predicted efficiency conditions (see below) to test one forecast against another. Forecast evaluation is an art in the same sense that forecasting itself is an art. It is reasonable to start testing miscalibration in several obvious directions, but to discover non-obvious miscalibration one has to be creative.

For a recursive one-step-ahead density forecasting situation an interesting question is whether UIPIT condition is sufficient for sequential auto-calibration. In general the answer is negative. However, when we are sure that the forecaster uses only the previous history to produce forecasts, then conditioning on the previous history of Y_t is equivalent to conditioning on the previous history of $P_t=\tilde{F}_t(Y_t)$ and, hence, UIPIT condition is sufficient not only for sequential auto-calibration, but for sequentially *ideal* calibration. That is, according to Theorem 16 under UIPIT each $\tilde{F}_t$ is ideally calibrated given $\sigma(Y_1,,\dots,Y_{t-1})$.

When forecasts are not measurable with respect to $\sigma(Y_1,\dots,Y_{t-1})$ UIPIT is not sufficient; it does not indicate a sequence of calibrated forecasts. Although for multistep forecasts, which are measurable with respect to $\sigma(Y_1,\dots,Y_{t-1})$, UIPIT condition is sufficient for sequentially ideal calibration, this is not a very useful sufficient condition since in general independence does not hold anyway. For forecasts using real-time data subject to revisions independence of PIT values can be completely irrelevant condition.

2.8 The general idea of moment-based calibration testing

As calibration tests in the existing literature mostly pertain to a situation when one-step-ahead forecasts of a time series are made given the full previous history of this series, these tests mostly rely on the UIPIT condition. Moreover, under UIPIT any functions of PIT values are also independent and have known distribution; for example, this is true of the tick (indicator) variables for interval/quantile forecasts. Therefore, under the UIPIT condition the distribution of the vector of observations is fully known, which facilitates construction of the corresponding tests. For example, likelihood ratio tests are widely used (e.g. Kupiec, 1995; Christoffersen, 1998; Berkowitz, 2001; Clements and Taylor, 2003).

In general we do not know the complete distribution of observations. The conditional distribution of a single Y_i given Ψ_i and $\tilde{F}$ is under the null of conditional auto-calibration with respect to $\tilde{F}$ fully described by the forecast $\tilde{F}_i$. However, to design tests we have to make assumptions on the dependence structure in the observations $i=1,\dots,N$.

Given a moment condition of calibration one can replace theoretical moments by sample ones based on a series of forecasts and outcomes and see how far the result is from what should be in theory. This allows to develop various types of diagnostic tests for forecast calibration. Chen (2011) demonstrates that many of the tests of calibration/efficiency developed in the literature fall within this approach.

Suppose that in theory the expectation of d must be zero under the null of calibration: $\mathsf{E}d=0$. We can obtain the values of d for a series of realizations of forecast functions $F_1,\dots,F_N$ and a series of outcomes $y_1,\dots,y_N$ and calculate the corresponding sample moment $\bar{d}=\sum_{i=1}^{N}d_i/N$.

If $\bar{d}$ is far from zero, then we can conclude that the forecast is miscalibrated.

Note that in order to test the moment conditions of calibration it is not necessary to assume that the data are described by some parametric model and that forecasts follow that model. Under appropriate assumptions on the distribution of the sequence of d_i, discussion of which is beyond the limits of this paper, we can use the usual t-ratios $\bar{d}/se(\bar{d})$. The most subtle aspect here is adequate calculation of the standard error $se(\bar{d})$ for dependent d_i. In the Example 3 below the usual heteroskedasticity and autocorrelation consistent (HAC) standard errors are used. If this is done correctly and the series of forecasts is well-calibrated, then this statistic is asymptotically distributed as $N(0,1)$. An extension to the multivariate case—simultaneous testing of several moment conditions—is straightforward and is familiar from the GMM framework: a t-ratio is replaced by a quadratic form and the distribution is chi-square. For the orthogonality conditions testing could be conveniently done by means of F-statistics and Wald statistics from auxiliary regressions (with robust covariance matrices if needed).

3 Forecast efficiency

3.1 Forecast efficiency, proper scoring rules and ideal forecasts

One can describe the simplest scheme of decision-making based on probabilistic forecasts as follows. The forecast user chooses some action a. The consequences depend on a realization y of a random variable Y. If the preferences of the forecast user are described by a utility function $u(a, y)$ and F is a realization of a probabilistic forecast of Y in the form of a distribution function, then the best action $a(F)$ is given by (e.g. Pesaran and Skouras, 2002)

$$a(F) \in \arg\max_a \mathsf{E}_{Y\sim F}[u(Y, a)].$$

(In forecasting theory one often uses expected loss minimization instead of expected utility maximization). One can say that $\tilde{F}_1$ is better than $\tilde{F}_2$ if it leads to a greater expected utility, that is,

$$\mathsf{E}u(Y, a(\tilde{F}_1)) > \mathsf{E}u(Y, a(\tilde{F}_2)).$$

This provides economic foundation for the theory of evaluation of probabilistic forecasts.

If we do not have a user with a utility function, then we can compare probabilistic forecasts using some suitable loss function or scoring rule.

A *scoring rule* is a function $S(F, y)$ of a CDF F and an outcome value y used to judge the accuracy or success of full probabilistic forecasts. If $F_1, \ldots, F_N$ is a series of realizations of predictive distribution functions, and $y_1, \ldots, y_N$ is a series of realized outcomes, then the average score is given by

$$\frac{1}{N}\sum_{i=1}^{N} S(F_i, y_i).$$

It is assumed that a more accurate (successful) forecast has a higher average score.

Not any arbitrary scoring rule is suitable for forecast evaluation. The general requirement is that scoring rules used for forecast evaluation must be *proper*.

One can define the expected score function as the expected score of F_2 given that Y is distributed as F_1:

$$S(F_2, F_1) = \mathsf{E}_{Y\sim F_1} S(F_2, Y).$$

(Note the overloaded notation used and take into account that both F_1 and F_2 are non-random CDFs here). By definition, if the scoring rule S is proper, then the expected score is maximized

with respect to F_2 when F_2 coincides with F_1:

$$S(F_1, F_1) \geq S(F_2, F_1),$$

and it is *strictly proper* (within a suitable class of distributions), if the inequality is strict for $F_2 \neq F_1$. Proper scoring rules are known to encourage truthful forecast statement: if the forecast is assessed according to a proper scoring rule, then the forecaster cannot expect to benefit by cheating and reporting forecast distributions which he believes to be incorrect.

A detailed review of this topic can be found in Gneiting and Raftery (2007) and Bröcker and Smith (2007). Economic applications of scoring rules can be found in Diebold and Rudebusch (1989), Clements and Harvey (2010), Boero, Smith, and Wallis (2011), Diks, Panchenko, and van Dijk (2011), Mitchell and Wallis (2011).

When forecasts are in the form of distribution functions it is logical to base forecast evaluation on the notion of a proper scoring rule, because it is closely related to the maximization of expected utility or minimizing expected loss by the forecast user. Indeed, define a scoring rule S as the utility of an outcome y under the best action $a(F)$:

$$S(F, y) = u(y, a(F)).$$

Such a utility-based scoring rule is proper since

$$S(F_1, F_1) = \mathsf{E}_{Y \sim F_1} u(Y, a(F_1)) \geq \mathsf{E}_{Y \sim F_1} u(Y, a(F_2)) = S(F_2, F_1)$$

(cf. Diebold, Gunther, and Tay, 1998, Gneiting and Raftery, 2007). Therefore, when analyzing the quality of probabilistic forecasts one can focus on proper scoring rules and abstract from the implicit expected utility maximization.

An important property of an ideally calibrated forecast is that it achieves the maximum expected score if the scoring rule used is proper. Diebold, Gunther, and Tay (1998), p. 866: "…If a forecast coincides with the true data generating process, then it will be preferred by all forecast users, regardless of loss function." See also Granger and Pesaran (2000).

According to Theorem 17 for any proper scoring rule the forecast, which is ideally calibrated with respect to Ψ, attains the highest expected score among the Ψ-measurable forecasts (Tsyplakov, 2011).

Thus, when a forecast is ideal with respect to Ψ, it can be called efficient or optimal. Under appropriate additional conditions the inequality in the theorem is strict if the scoring rule S is strictly proper and the alternative forecast is not ideal (Holzmann and Eulert, 2011).

Therefore, if a forecast is not auto-calibrated given Ψ, which is signaled by a violation of some moment condition, then it is not ideally calibrated given Ψ and $\tilde{F}$ and there is a potential for its improvement with the help of the information contained in Ψ and $\tilde{F}$. Improvement is measured by an increase in the mean score.

On the other hand, if Ψ^* is the information set containing all available information and $\Psi \subset \Psi^*$, then the forecast which is ideally calibrated given Ψ^* and hence has the largest expected score is auto-calibrated given Ψ. This means that an efficient forecast would not be dismissed by the auto-calibration criterion.

It can be said that in a certain sense the concept of calibration is intrinsically based on proper scoring rules and score maximization.

It is also notable that each general moment condition (1) suggests the corresponding proper scoring rule

$$S_g(F, y; w) = -(g(y, w, F) - \gamma(F, w))^2.$$

3.2 Moment conditions of forecast efficiency

We can use the first order conditions of score maximization to derive moment conditions of efficiency. Consider a CDF-to-CDF transformation $T(F, w, \delta)$ depending on a real vector of parameters δ and an additional variable w. We require that $F = T(F, w, 0)$. Suppose that $\tilde{F}$ is an efficient forecast, Ψ is the relevant information set and W is some Ψ-measurable random element. The transformation T can produce a family of forecasts $\tilde{F}_\delta = T(\tilde{F}, W, \delta)$ parametrized by δ, which includes the efficient forecast $\tilde{F}$ with $\delta = 0$. If $\mathsf{E}S(\tilde{F}_\delta, Y)$ is differentiable as a function of δ, then we must have

$$\frac{d}{d\delta}\mathsf{E}S(\tilde{F}_\delta, Y)\Big|_{\delta=0} = 0.$$

Under appropriate regularity conditions the differentiation and expectation operations are interchangeable and we obtain the following moment conditions:

$$\mathsf{E}\frac{d}{d\delta}S(\tilde{F}_\delta, Y)\Big|_{\delta=0} = 0. \tag{10}$$

The idea here is that we can extend a forecast $\tilde{F}$ in a parametric way (irrespectively of a possible parametric model on which $\tilde{F}$ could be based) and then derive moment conditions which follow from the fact that the maximum score is attained for those parameters which correspond to the initial forecast $\tilde{F}$ if $\tilde{F}$ is efficient.

By the same logic assuming that S is proper we must have for an arbitrary CDF F

$$\mathsf{E}_{Y\sim F}\frac{d}{d\delta}S(F_\delta, Y)\Big|_{\delta=0} = 0,$$

where $F_\delta = T(F, w, \delta)$. It can be seen that efficiency conditions (10) can be considered as autocalibration conditions given Ψ of general type (1) with $g = \frac{d}{d\delta}S(T(F, w, \delta), y)\,|_{\delta=0}$.

Location A simple transformation of a CDF is a shift by $w^T\delta$ where w is a real vector (which may include a constant element 1):

$$F_\delta(y) = F(y - w^T\delta).$$

For example, consider a density forecast with log-density

$$\tilde{\ell}(y) = \log\tilde{F}'(y)$$

and the logarithmic scoring rule

$$S(F, y) = \log F'(y).$$

In this case (necessary) moment conditions of forecast efficiency are given by

$$\mathsf{E}[-\tilde{\ell}'(Y)W] = 0,$$

where W is a Ψ-measurable vector.

Scale Another simple transformation is scaling of CDF F around some central point $c(F)$. Natural central points are the median $c = F^{-1}(1/2)$ and the mean $c = \mathrm{mean}(F)$:

$$F_\delta(y) = F((y - c(F))\exp(-w^T\delta) + c(F)).$$

For the logarithmic scoring rule the corresponding conditions of forecast efficiency are given by

$$\mathsf{E}[(-\tilde{\ell}'(Y)(Y - c(\tilde{F})) - 1)W] = 0.$$

Inverse normal transform: location and scale Alternatively, we can employ transformations based on the inverse normal transform (INT) of CDF $\tilde{F}$ defined as $\Phi^{-1} \circ F$, where $\Phi(\cdot)$ is the standard normal CDF:

$$F_\delta(y) = \Phi(\Phi^{-1}(F(y)) - w^T\delta)$$

and

$$F_\delta(y) = \Phi(\Phi^{-1}(F(y)) \exp(-w^T\delta)).$$

These transformations correspond to the location and scale and give the following conditions of forecast efficiency with the logarithmic scoring rule:

$$\mathsf{E}[\mathrm{INT}W] = 0$$

and

$$\mathsf{E}[(\mathrm{INT}^2 - 1)W] = 0,$$

where $\mathrm{INT} = \Phi^{-1}(\tilde{F}(Y))$. It can be seen that the two conditions are orthogonality conditions for probabilistic calibration of type (5). This demonstrates that some known calibration tests based on PIT and INT values (e.g. Berkowitz, 2001) can be motivated by their connection with efficiency.

3.3 Calibration, efficiency and sharpness

Another link between calibration and efficiency is provided by the sharpness principle of probabilistic forecasting.

Forecast sharpness is a characteristic which reflects the degree of forecast definiteness, the concentration of the forecast distribution (Murphy and Winkler, 1987; Gneiting, Balabdaoui, and Raftery, 2007). Users can prefer sharp forecast as they are more definite and informative. However, forecast sharpness can be deceptive and it is not a good idea to make choice between forecasts solely on the basis of their sharpness.

In Gneiting, Balabdaoui, and Raftery (2007) a conjecture called "the sharpness principle" was put forward, which states that the problem of finding a good forecast can be viewed as the problem of maximizing sharpness subject to calibration. It can be shown that the conjecture is actually true provided that a vague "calibration" notion is replaced by (conditional or unconditional) auto-calibration.

First, for a proper scoring rule $S(F, F)$ can be viewed as a measure of sharpness of a distribution F. For a proper scoring rule $-S(F, F)$ is a concave[10] function of F and thus, according to DeGroot (1962), can be viewed as a measure of uncertainty of a probability distribution F. For the logarithmic scoring rule $-S(F, F)$ is the familiar Shannon's entropy measure.

Second, for an auto-calibrated forecast we have $\mathsf{E}S(\tilde{F}, Y) = \mathsf{E}S(\tilde{F}, \tilde{F})$, i.e. the expected score of such a forecast equals its expected sharpness. The fact follows from (1) for $g = S(F, y)$.

This means that auto-calibrated forecasts can be compared on the basis of the levels of their expected sharpness. Sharpness is no more a deceptive characteristic when only (unconditionally) auto-calibrated forecasts are considered. The ideally calibrated forecast given Ψ is the sharpest of all Ψ-measurable auto-calibrated forecasts, because it is characterized by the greatest expected score.

Another intuitively expected property of well-calibrated forecasts is that the more complete information has the forecaster, the sharper is the forecast which he can potentially produce. Let

[10]Function $S(F_1, F_2)$ is linear in the second argument. Therefore $S(F_\alpha, F_\alpha) = \alpha S(F_\alpha, F_1) + (1-\alpha)S(F_\alpha, F_2) \leq \alpha S(F_1, F_1) + (1-\alpha)S(F_2, F_2)$ for $F_\alpha = \alpha F_1 + (1-\alpha)F_2$ and $\alpha \in [0, 1]$.

$\mathbb{F}_1 = \mathbb{F}_{\Psi_1}$ be the ideal forecast based on Ψ_1 and $\mathbb{F}_2 = \mathbb{F}_{\Psi_2}$ the ideal forecast based on Ψ_2, where Ψ_1 is a "richer" information set than Ψ_2 ($\Psi_2 \subset \Psi_1$). Then

$$\mathsf{E}S(\mathbb{F}_1, Y) = \mathsf{E}S(\mathbb{F}_1, \mathbb{F}_1) \geq \mathsf{E}S(\mathbb{F}_2, Y) = \mathsf{E}S(\mathbb{F}_2, \mathbb{F}_2)$$

with strict inequality if $\mathbb{F}_1 \neq \mathbb{F}_2$ almost surely and S is strictly proper. See Holzmann and Eulert (2011) for a proof. Similar results for the discrete outcome case can be found in DeGroot and Fienberg (1983) and Bröcker (2009).

We can further study the relationship between the expected score and the expected sharpness for forecasts, which lack conditional auto-calibration. Let d denote a divergence indicator (generalized distance) between distributions F_1 and F_2 defined as

$$d(F_2, F_1) = S(F_1, F_1) - S(F_2, F_1).$$

The divergence $d(F_2, F_1)$ is non-negative, if the rule S is proper. It is zero when the two distributions coincide. For the logarithmic scoring rule d is the Kullback–Leibler distance. In general since $\mathsf{E}S(\tilde{F}, Y) = \mathsf{E}S(\tilde{F}, \mathbb{F}_{\Psi,\tilde{F}})$ the expected score of a (possibly miscalibrated) forecast $\tilde{F}$ can be decomposed as follows:

$$\mathsf{E}S(\tilde{F}, Y) = \mathsf{E}S(\mathbb{F}_{\Psi,\tilde{F}}, \mathbb{F}_{\Psi,\tilde{F}}) - \mathsf{E}d(\tilde{F}, \mathbb{F}_{\Psi,\tilde{F}}),$$

where $\mathbb{F}_{\Psi,\tilde{F}}$ is the conditional distribution function of Y given Ψ and $\tilde{F}$. The first term can be interpreted as the expected sharpness of the forecast $\mathbb{F}_{\Psi,\tilde{F}}$, which is a "fully recalibrated" version of forecast $\tilde{F}$ given Ψ, while the second term relates to the divergence between $\tilde{F}$ and $\mathbb{F}_{\Psi,\tilde{F}}$, i.e. it is a measure of miscalibration of forecast $\tilde{F}$ with respect to the information contained in itself and Ψ. A version of this partitioning for the dichotomous outcomes and the Brier score was developed in Sanders (1963). Bröcker (2009) extended it to the case of an arbitrary finite-support discrete distribution and arbitrary proper scoring rules.

The principle of maximizing sharpness subject to calibration which was considered here is difficult to apply in practice, because achieving perfect conditional auto-calibration of a forecast may prove too challenging. However, this principle provides a useful insight into the essence of probabilistic forecasting. In particular, it is clear that the advantage of using proper scoring rules for forecast comparison is that they provide the right balance of sharpness and calibration. If other—not proper—scoring rules were used for forecast evaluation, then the forecaster would have an incentive to report miscalibrated (for example, too sharp) forecasts.

3.4 Predicted efficiency conditions

Finally in this section we consider calibration conditions, which relate to forecast efficiency indirectly, through its use of proper scoring rules.

As was already noted above, the expected sharpness of an auto-calibrated forecast equals its expected score. I general if $\tilde{F}$ is auto-calibrated given Ψ, then from (4) with $r = S(F, y)$ we have that for any function $a(w, F)$ and any Ψ-measurable W

$$\mathsf{E}[(S(\tilde{F}, Y) - S(\tilde{F}, \tilde{F}))\, a(W, \tilde{F})] = 0. \tag{11}$$

Another condition is a variety of forecast encompassing condition. Define g as the score differential between F_1 and F_2[11]

$$g = S(F_2, y) - S(F_1, y).$$

[11] Logarithmic score differential in the form of likelihood ratio or Kullback–Leibler information criterion is used for forecast evaluation purposes in Amisano and Giacomini (2007) and Bao, Lee, and Saltoğlu (2007); see also Hall and Mitchell (2007).

and let γ be the expectation of g under assumption that Y is distributed as F_1

$$\gamma = S(F_2, F_1) - S(F_1, F_1).$$

According to the general moment conditions of calibration (1) for two forecasts $\tilde{F}_1$, $\tilde{F}_2$ under the assumption that $\tilde{F}_1$ is auto-calibrated with respect to $\Psi \cup \sigma(\tilde{F}_2)$ we have

$$\mathsf{E}[S(\tilde{F}_2, Y) - S(\tilde{F}_1, Y)] = \mathsf{E}[S(\tilde{F}_2, \tilde{F}_1) - S(\tilde{F}_1, \tilde{F}_1)]. \tag{12}$$

This moment condition can be called a *relative predicted efficiency* (RPE) condition[12]. The relative predicted efficiency conditions parallel the generalization of the likelihood ratio test for non-nested models developed in Cox (1961), Cox (1962).

4 Examples

4.1 Example 1, the pitfalls of the UIPIT condition

Consider the following artificial example. Starting from $Y_0 \sim N(0,1)$ define Y_t for $t \geq 1$ recursively:

$$\begin{gathered} Y_t = \mu_t + \epsilon_t, \\ \mu_t = \Phi^{-1}(\{K\Phi(Y_{t-1})\})\sqrt{1-\lambda}, \\ \epsilon_t | Y_0, \ldots, Y_{t-1} \sim N(0, \lambda), \end{gathered}$$

where $\Phi(\cdot)$ is the standard normal CDF, $\{\cdot\}$ is the fractional part, K is a large integer and $\lambda \in (0,1)$.

We assume that the relevant information set for the forecast at time $t \geq 1$ is $\Psi_t = \sigma(Y_0, Y_1, \ldots, Y_{t-1})$. Three different forecasts are considered.

Forecast A: $N(\mu_t, \lambda)$. This forecast reflects the DGP and is ideally calibrated with respect to Ψ_t.

Forecast B: $N(0,1)$. This forecast corresponds to the unconditional distribution of Y_t and is unconditionally auto-calibrated. PIT values of this forecast are dependent. However, the DGP incorporates a highly non-linear transformation, which disguises the dependence. For large K it is impossible to find any traces of serial dependence in the series of the PIT values for Forecast B by the means of PIT-based tests ordinarily used for forecast evaluation.

Forecast C: $N(\mu_t + \xi_t, \lambda + \beta)$, where ξ_t is an independent Gaussian white noise $\xi_t \sim N(0, \beta)$. Forecast C is based on Forecast A, but contains additional noise. It has PIT values $\Phi((\epsilon_t - \xi_t)/\sqrt{\lambda+\beta})$, which are distributed as $U[0,1]$ and independent. Moreover, PIT values are distributed as $U[0,1]$ conditionally on Ψ_t (that is, probabilistically calibrated given Ψ_t). However, Forecast C is not auto-calibrated with respect to Ψ_t, since it is not marginally calibrated given Ψ_t.

Here we have two forecasts with uniform and independent PIT values and one forecast with uniform PIT values and non-obvious dependence in PIT values. We can run a battery of tests for PIT uniformity and independence such as those listed in Chen (2011) and Mitchell and Wallis (2011). However, the result of such exercise is foreseeable so we skip it.

[12] In Tsyplakov (2011) it was called relative forecast calibration condition.

Table 1: Statistics for the forecasts of Example 2

	$\tilde{F}_c$	$\tilde{F}_{r1}$	$\tilde{F}_{r2}$	$\tilde{F}_{xz}$
Expected log. score	−1.612	−1.596	−1.525	−1.484
% best	0	0	1.41	98.59
Test 1, INT$\perp 1, X$	99.95	99.95	4.22	4.99
Test 2, $Y - M \perp 1, X$	99.97	99.97	5.15	5.09
Test 3, INT$^2 - 1 \perp 1$	83.54	5.84	5.92	5.99
Test 4, PE	98.41	60.22	61.52	5.84
Test 5, RPE	99.99	99.57	11.98	4.57

Note: The table is based on 10000 simulations. The expected logarithmic score is in the first row. The test statistics are F-statistics. The figures for the tests are rejection frequencies are for 5% asymptotic significance level using the corresponding F quantiles.

The example is artificial and is not directly related to real forecasting problems, but it is suggestive. Although from the point of view of the usual tests for the UIPIT condition all the three forecasts look indistinguishably perfect, they are different in terms of efficiency. For example, the expected logarithmic score of a single forecast of Y_t is given by $K - \log(\lambda)/2$ for Forecast A, K for Forecast B and $K - \log(\lambda + \beta)/2$ for Forecast B, where $K = -(\log(2\pi) + 1)/2$.

Forecast A is dramatically better than B for $\lambda \ll 1$ and dramatically better than C for $\beta \gg \lambda$. Comparison of Forecasts A and B shows that it is not advisable to rely on direct tests of the UIPIT condition when the time series can incorporate non-obvious non-linearity. Comparison of Forecasts A and C demonstrates a possible problem with reliance on the UIPIT condition and conditional probabilistic calibration when the forecast can include extraneous noise.

4.2 Example 2, combined forecasts, simulation

Our second example relates to calibration testing of combined forecasts. Suppose that Y is given by

$$Y = X + \epsilon/\sqrt{Z},$$

where $X \sim N(0,1)$, $Z \sim \Gamma_{1/2,2}$ and $\epsilon \sim N(0,1)$ are independent. Also denote $\tilde{F}_x$, $\tilde{F}_z$, $\tilde{F}_{xz}$ conditional CDFs which correspond to $Y|X \sim t_8$, $Y|Z \sim N(0, 1 + 1/Z)$ and $Y|X, Z \sim N(X, 1/Z)$.

We run simulations for four forecasts. The first is the equal-weight linear pool of two partial conditional CDFs: $\tilde{F}_c = \frac{1}{2}\tilde{F}_x + \frac{1}{2}\tilde{F}_z$. The second and third forecasts are recalibrated versions of $\tilde{F}_c$. The recalibration is implemented via an INT-based model:

$$\mathrm{INT}_c = \beta X + \xi, \qquad \mathsf{Var}\xi = \sigma,$$

where $\mathrm{INT}_c = \Phi^{-1}(\tilde{F}_c(Y))$ is the inverse normal transform for $\tilde{F}_c$. The recalibrated forecast is given by $\Phi((\Phi^{-1}(\tilde{F}_c(Y)) - \beta X)/\sigma)$. Forecast $\tilde{F}_{r1}$ with $\beta = 0$, $\sigma = 0.874$ repairs only the incorrect unconditional dispersiveness of $\tilde{F}_c$. Forecast $\tilde{F}_{r2}$ with $\beta = 0.316$, $\sigma = 0.814$ also repairs the conditional mean. (The parameters are approximations to the corresponding theoretical models.) Finally, forecast $\tilde{F}_{xz}$ is known to be conditionally auto-calibrated with respect to $\sigma(X)$ and can be regarded as a perfectly recalibrated variant of $\tilde{F}_c$.[13].

We reproduce a situation where a forecast examiner can observe X, but not Z or ϵ. Some forecaster(s) presented him forecasts $\tilde{F}_c$, $\tilde{F}_{r1}$ and $\tilde{F}_{r2}$. (We are adding $\tilde{F}_{xz}$ for control purposes.)

[13]Note that $\sigma(\tilde{F}_c) = \sigma(X, Z)$.

From his point of view the suitable mode of calibration is conditional auto-calibration with respect to $\Psi = \sigma(X)$. Five different tests are used, which are based on the following moment conditions.

Test 1 $\mathsf{E}\mathrm{INT} = 0$ and $\mathsf{E}[\mathrm{INT}X] = 0$, where $\mathrm{INT} = \Phi^{-1}(\tilde{F}(Y))$.

Test 2 $\mathsf{E}[Y - M] = 0$ and $\mathsf{E}[(Y - M)X] = 0$, where $M = \mathrm{mean}(\tilde{F})$.

Test 3 $\mathsf{E}[\mathrm{INT}^2 - 1] = 0$.

Test 4 Predicted efficiency test $\mathsf{E}S_{log}(\tilde{F}, Y) = \mathsf{E}S_{log}(\tilde{F}, \tilde{F})$ and $\mathsf{E}[(S_{log}(\tilde{F}, Y) - S_{log}(\tilde{F}, \tilde{F}))S_{log}(\tilde{F}, \tilde{F})] = 0$, where $S_{log}(F, y) = \log F'(y)$ is the logarithmic scoring rule.

Test 5 RPE test against $\tilde{F}_x$ based on $\mathsf{E}[S_{log}(\tilde{F}_x, Y) - S_{log}(\tilde{F}, Y)] = \mathsf{E}[S_{log}(\tilde{F}_x, \tilde{F}) - S_{log}(\tilde{F}, \tilde{F})]$.

We used simulations with 200 observations. The required moments in intractable cases were calculated by plain numerical integration. Table 1 shows the results on the expected logarithmic score, comparison of average logarithmic scores and rejection rates for five calibration tests.

The expected logarithmic score shows the asymptotic potential of a forecast which becomes visible when the number of observations tends to infinity. When a series of forecasts is not very long, imperfect forecasts can obtain higher average scores than the ideal forecast. So "% best" row shows the percentage of experiments in which the corresponding model had the highest average logarithmic score.

The basic combined forecast $\tilde{F}_c$ is underdispersed and $\tilde{F}_{r1}$ corrects this well-known problem (cf. Gneiting and Ranjan, 2013 where the beta CDF is used for the same purpose). Test 3, indeed, frequently signals inadequate unconditional dispersiveness in $\tilde{F}_c$, while it shows rejection rate close to 5% for the three recalibrated forecasts.

Recalibration of $\tilde{F}_c$ for both location and scale produces forecast $\tilde{F}_{r2}$, which does not show obvious signs of either probabilistic or marginal conditional miscalibration. However, since it does not coincide with the ideally recalibrated forecast $\tilde{F}_{xz}$, it cannot be auto-calibrated. Indeed, Test 4 often signals miscalibration. Interestingly, Test 5 (RPE test) also has some power against this miscalibration in $\tilde{F}_{r2}$.

Here partial miscalibration criteria (like conditional probabilistic calibration) are not able to signal miscalibration in $\tilde{F}_{r2}$ conditionally on X. Even though Z is unobservable to the examiner directly, he can use the characteristics of the forecast itself to detect miscalibration. It can be also seen from this example that an RPE test can be useful in situations where the direction of miscalibration is not obvious. If there exists some rough forecast then we can use it for miscalibration diagnostics without additional analysis. Tests 4 and 5 use general moment conditions (1) of forecast calibration and they cannot be reduced to testing conditional probabilistic or marginal calibration given X. Thus the concept of conditional auto-calibration can be important in practice, not only in theory.

4.3 Example 3, stock index forecasts

As our third example we consider one-step-ahead forecasting of the daily returns of a stock index. The data are daily returns (in percent) $R_t = (\log RTSI_t - \log RTSI_{t-1}) \times 100$ of the Russian stock market index (RTSI) calculated from the close levels. We use the data for the period from 1995-09-01 to 2013-08-09 (4,478 observations).

The following forecasts are considered.

Table 2: Statistics for the three forecasts of RTSI, Example 3

	200-t	ES	AR-GARCH-t
Log. score	−2.217	−2.202	−2.148
Location test 1	−1.11	2.80**	−1.07
Location test 2	5.54***	8.24***	−1.92
Scale test 1	0.30	2.29*	−3.16**
Scale test 2 vs. 200-t	−1.17	1.43	−1.44
Scale test 2 vs. ES	2.72**	1.03	−1.76
Scale test 2 vs. AR-GARCH-t	3.80***	2.14*	−3.04**

Note: The figures for the tests are t-ratios. Newey–West HAC standard errors with lag truncation 3 where used. Statistical significance at 5% (1%, 0.1%) level is shown by * (**, ***).

Forecast 200-t: A rolling span of the 200 most recent observations $y_{t-199}, \ldots, y_t$ is used to fit the Student's t distribution with location parameter α, scale parameter σ and degrees-of-freedom parameter ν by the maximum likelihood.

Forecast ES: The forecast is based on exponential smoothing for volatility $\sigma_{t+1}^2 = (1-\delta)R_t^2 + \delta\sigma_t^2$ with the decay factor $\delta = 0.95$ (RiskMetrics, 1996). The one-step-ahead forecasting distribution is given by $N(0, \sigma_{t+1}^2)$. The recursion for volatility starts from the sample variance of the first 200 observations.

Forecast AR-GARCH-t: The forecast is based on AR(1)-GARCH(1,1)-t model (the first order autoregression, where disturbances are GARCH(1,1)-t as in Bollerslev, 1987). The model is estimated recursively by the maximum likelihood method. The forecasting distribution is Student's t with location, scale and degrees-of-freedom parameters supplied by the model.

All the forecasts are produced starting from the observation 201. They are compared by their observed average logarithmic scores. The following statistics are summarized in Table 2.

Log. score is the average logarithmic score.

Location test 1 is the test corresponding to the moment condition $\mathsf{E}[-\tilde{\ell}'(Y)] = 0$. For the Student's t distribution with parameters α, σ and ν the derivative of the log-density used in the test is given by

$$-(\nu+1)(y-\alpha)/(\nu\sigma^2 + (y-\alpha)^2).$$

Location test 2 corresponds to the moment condition $\mathsf{E}[-\tilde{\ell}'(Y)\,Y_{-1}] = 0$, where Y_{-1} is the first lag of Y.

Scale test 1 corresponds to the moment condition $\mathsf{E}[-\tilde{\ell}'(Y)(Y-C)] = 1$, where C is the center of the predictive distribution (which is defined unambiguously for symmetric forecasts such as used here).

Scale test 2 vs. $\tilde{H}$ corresponds to the moment condition $\mathsf{E}[(-\tilde{\ell}'(Y)(Y-C) - 1)R] = 0$, where $R = \log(\tilde{H}^{-1}(3/4) - \tilde{H}^{-1}(1/4))$ is a scale variable defined as the logarithm of the interquartile range of the rival forecast $\tilde{H}$.

Forecast AR-GARCH-t is expectedly the best according to the average logarithmic score. ES is the only forecast which does not include a constant term and Location test 1 signals the resulting downward bias.

AR-GARCH-t includes an autoregressive term and is the only forecast which is not rejected by Location test 2.

Scale test 1 shows that ES produces too narrow forecasts while the forecasts from AR-GARCH-t are too wide. Scale test 2 against AR-GARCH-t confirms this lack of calibration.

Although 200-t has unconditionally correct scale, Scale test 2 against ES and AR-GARCH-t demonstrates that volatility clustering is not fully captured.

We can conclude that all three forecasts are not fully efficient. The efficiency tests show the directions of possible improvement.

5 Conclusion

In evaluation of probabilistic forecasts it is desirable to rely on fundamental concepts and theoretical properties. Some of such concepts and properties were considered in this paper.

Conditional auto-calibration and ideal calibration given an information set are two important fundamental concepts which can be used.

Among other things, the concept of auto-calibration helps to derive the principle of maximizing sharpness subject to calibration from expected score maximization.

The paper highlights the difference between conditional auto-calibration and the less general concepts of conditional probabilistic calibration (PIT uniformity) and conditional marginal calibration.

The paper argues that expected score maximization and the notion of a proper scoring rule can be viewed as the implicit basis for evaluation of probabilistic forecasts. The notion of calibration can be derived from this basis.

Forecast efficiency, conditional marginal and probabilistic calibration, conditional auto-calibration can be expressed by various moment conditions, including orthogonality conditions. These conditions lead to a general framework for calibration testing. The framework can facilitate construction of various new tests. This is exemplified by general forecast encompassing tests introduced in this paper, including tests based on relative predicted efficiency condition.

Last, but not the least, the paper suggests that a great caution is required when using the condition of uniformity and independence of PIT values as a definition of ideal calibration or efficiency. The moment conditions described here can be used to extend forecast evaluation techniques to situations where this condition is not necessary or inappropriate.

Appendix

Theorem 6. *Consider a measurable real-valued function $b(x, y)$ and two random elements, X and real-valued Y, such that $\mathsf{E}|b(X, Y)| < \infty$. If X is measurable with respect to a sub-σ-algebra Ψ, then*

$$\mathsf{E}_\Psi b(X, Y) = \beta(X, \mathbb{F}_\Psi)$$

and

$$\mathsf{E} b(X, Y) = \mathsf{E}\beta(X, \mathbb{F}_\Psi),$$

where $\mathbb{F}_\Psi(y)$ is the conditional CDF of Y given Ψ and

$$\beta(x,F) = \mathsf{E}_{Y\sim F} b(x,Y) = \int b(x,y)\,dF(y).$$

This is a corollary of Theorem 6.4 (*disintegration theorem*) in Kallenberg (2002), p. 108. Theorem 6 can be viewed as a combination of two properties of conditional expectations and conditional distributions. First, conditional expectations can be represented by the unconditional expectations in terms of the corresponding conditional distributions:

$$\mathsf{E}_\Psi m(Y) = \mu(\mathbb{F}_\Psi),$$

where

$$\mu(F) = \mathsf{E}_{Y\sim F} m(Y) = \int m(y)\,dF(y).$$

Second, if a random element X is Ψ-measurable, then it can be treated as fixed inside the conditional expectation with respect to Ψ (the substitution property of the conditional expectation):

$$\mathsf{E}_\Psi b(X,Y) = B(X),$$

where $B(x) = \mathsf{E}_\Psi b(x,Y) = \beta(x,\mathbb{F}_\Psi)$. For $b(x,y) = xy$ the substitution property is a generalization of the well-known property of conditional distribution: $\mathsf{E}_\Psi[XY] = X\mathsf{E}_\Psi Y$, if X is Ψ-measurable.

Theorem 7. *If a forecast $\tilde{F}$ is conditionally auto-calibrated with respect to Ψ then for any g and any Ψ-measurable W it satisfies*

$$\mathsf{E}_{\Psi,\tilde{F}} g(Y,W,\tilde{F})| = \gamma(\tilde{F},W)$$

and

$$\mathsf{E} g(Y,W,\tilde{F}) = \mathsf{E}\gamma(\tilde{F},W),$$

where

$$\gamma(F,w) = \mathsf{E}_{Y\sim F} g(Y,w,F).$$

Proof. Denote $\gamma_0(H,w,F) = \mathsf{E}_{Y\sim H} g(Y,w,F)$, where F,H are non-random distribution functions. Since $\tilde{F}$ and W are measurable with respect to $\Psi\cup\sigma(\tilde{F})$, then by Theorem 6

$$\mathsf{E}_{\Psi,\tilde{F}} g(Y,W,\tilde{F}) = \gamma_0(\mathbb{F}_{\Psi,\tilde{F}},W,\tilde{F}).$$

If $\tilde{F} = \mathbb{F}_{\Psi,\tilde{F}}$, then the right-hand side is $\gamma_0(\tilde{F},W,\tilde{F}) = \gamma(\tilde{F},W)$. Then the law of iterated expectations can be applied. □

Theorem 8. *If a forecast $\tilde{F}$ is conditionally auto-calibrated with respect to Ψ, then it is marginally calibrated given Ψ and (in a density forecasting situation) probabilistically calibrated given Ψ.*

Proof. [conditional marginal calibration] Note that $\mathsf{E}_\Psi\tilde{F}(y) = \mathsf{E}_\Psi\mathbb{F}_{\Psi,\tilde{F}}(y) = \mathbb{F}_\Psi(y)$ for any real y.

[conditional probabilistic calibration] According to Theorem 7 under conditional auto-calibration with respect to Ψ for $g = \mathrm{I}\{F(y)\le p\}$ and $\gamma = p$ we have $\mathsf{E}_\Psi\mathrm{I}\{\tilde{F}(Y)\le p\} = p$. □

Theorem 9. *Probabilistic calibration given Ψ is equivalent to the condition that (1) $P\sim U[0,1]$ and (2) P and Ψ are independent.*

Proof. Let $A = P^{-1}(-\infty, p]$ for some p and $B \in \Psi$. Then

$$\mathsf{P}(A \cap B) = \mathsf{E}[\mathrm{I}\{P \le p\}\mathrm{I}(B)] = \mathsf{E}\mathsf{E}_\Psi[\mathrm{I}\{P \le p\}\mathrm{I}(B)] = \mathsf{E}[\mathsf{E}_\Psi \mathrm{I}\{P \le p\} \cdot \mathrm{I}(B)] = p\mathsf{E}\mathrm{I}(B) = \mathsf{P}(A)\mathsf{P}(B).$$

It follows that P and Ψ are independent.

Conversely, suppose that $P \sim U[0,1]$ and that P and Ψ are independent. Since $\mathrm{I}\{P \le p\}$ and Ψ are independent we have $\mathsf{E}_\Psi \mathrm{I}\{P \le p\} = \mathsf{E}\mathrm{I}\{P \le p\} = p$. □

Theorem 10. *If a forecast $\tilde{F}$ is marginally calibrated given Ψ (that is, $\mathsf{E}_\Psi \tilde{F} = \mathbb{F}_\Psi$) then for any n and any Ψ-measurable W it satisfies*

$$\mathsf{E}n(Y, W) = \mathsf{E}v(\tilde{F}, W),$$

where

$$v(F, w) = \mathsf{E}_{Y \sim F} n(Y, w).$$

Proof. By Theorem 6 we have $\mathsf{E}_\Psi n(Y, W) = v(\mathbb{F}_\Psi, W)$. Since $v(F, w)$ is linear in F, it follows that

$$v(\mathbb{F}_\Psi, W) = v(\mathsf{E}_\Psi \tilde{F}, W) = \mathsf{E}_\Psi v(\tilde{F}, W).$$

Then the law of iterated expectations can be applied. □

Theorem 11. *If a forecast $\tilde{F}$ is ideal with respect to Ψ^*, then it is conditionally auto-calibrated with respect to Ψ for any $\Psi \subseteq \Psi^*$.*

Proof. Since $\Psi \cup \sigma(\tilde{F}) \subset \Psi^*$ we have $\mathbb{F}_{\Psi,\tilde{F}} = \mathsf{E}_{\Psi,\tilde{F}} \mathbb{F}_{\Psi^*} = \mathsf{E}_{\Psi,\tilde{F}} \tilde{F} = \tilde{F}$. □

Theorem 12. *Suppose that in a recursive one-step density forecasting situation each forecast $\tilde{F}_t$, $t = 1, \dots, T$ is auto-calibrated with respect to $\Psi_t = \sigma(Y_1, \dots, Y_{t-1}, \tilde{F}_1, \dots, \tilde{F}_{t-1})$.*[14] *Then $(P_1, \dots, P_T) \sim U[0,1]^T$, where $P_t = \tilde{F}_t(Y_t)$.*

Proof. By Theorem 8 we have $P_t | \Psi_t \sim U[0,1]$. Since $\sigma(P_1, \dots, P_{t-1}) \subset \Psi_t$, it follows that $P_t | P_1, \dots, P_{t-1} \sim U[0,1]$. Using this fact and starting induction from $P_1 \sim U[0,1]$ we obtain $(P_1, \dots, P_T) \sim U[0,1]^T$. □

Example 13. The actual distribution of Y is described by $\mathbb{F}(y|W) = F^\circ_W(y)$, where $W = 1$ or $W = 2$ with equal probabilities and

$$F^\circ_1(y) = \begin{cases} \frac{3}{2}y, & y \in [0, \frac{1}{4}], \\ \frac{1}{2}y + \frac{1}{4}, & y \in [\frac{1}{4}, \frac{3}{4}], \\ \frac{3}{2}y - \frac{1}{2}, & y \in [\frac{3}{4}, 1], \end{cases} \qquad F^\circ_2(y) = \begin{cases} \frac{1}{2}y, & y \in [0, \frac{1}{4}], \\ \frac{3}{2}y - \frac{1}{4}, & y \in [\frac{1}{4}, \frac{3}{4}], \\ \frac{1}{2}y + \frac{1}{2}, & y \in [\frac{3}{4}, 1]. \end{cases}$$

The forecast $\tilde{F}$ is also based on W: $\tilde{F}(y) = F_W(y)$, where

$$F_1(y) = \begin{cases} y, & y \in [0, \frac{1}{2}], \\ \frac{1}{2}y + \frac{1}{4}, & y \in [\frac{1}{2}, \frac{3}{4}], \\ \frac{3}{2}y - \frac{1}{2}, & y \in [\frac{3}{4}, 1], \end{cases} \qquad F_2(y) = \begin{cases} y, & y \in [0, \frac{1}{2}], \\ \frac{3}{2}y - \frac{1}{4}, & y \in [\frac{1}{2}, \frac{3}{4}], \\ \frac{1}{2}y + \frac{1}{2}, & y \in [\frac{3}{4}, 1]. \end{cases}$$

Since $\frac{1}{2}F^\circ_1(y) + \frac{1}{2}F^\circ_2(y) = \frac{1}{2}F_1(y) + \frac{1}{2}F_2(y)$ ($= y$ for $y \in [0,1]$) and $\frac{1}{2}F^\circ_1(F_1^{-1}(p)) + \frac{1}{2}F^\circ_2(F_2^{-1}(p)) = p$, it can be seen that $\tilde{F}$ is both marginally and probabilistically calibrated. However, $\sigma(\tilde{F}) = \sigma(W)$ and thus for $\tilde{F}$ to be auto-calibrated we must have $\tilde{F} = F^\circ_W$ which is not the case here.

[14] Here Ψ_1 is assumed to be the trivial σ-algebra.

Theorem 14. *If in a density forecasting situation for a Ψ-measurable forecast $\tilde{F}$ we have $\tilde{F}(Y)|\Psi \sim U[0,1]$ then $\tilde{F}$ is ideal with respect to Ψ.*

Proof. $\mathbb{F}_\Psi(y) = \mathsf{E}_\Psi \mathrm{I}\{Y \le y\} = \mathsf{E}_\Psi \mathrm{I}\{\tilde{F}(Y) \le \tilde{F}(y)\} = \tilde{F}(y)$. □

Theorem 15. *If a forecast $\tilde{F}$ satisfies condition*

$$\mathsf{E}_\Psi[r(Y,\tilde{F}) - \rho(\tilde{F})] = 0,$$

for any r, where

$$\rho(F) = \mathsf{E}_{Y\sim F} r(Y,F),$$

then it is conditionally auto-calibrated with respect to Ψ.

Proof. Let $r(y,F) = \mathrm{I}\{y \le y_0\}a(w,F)$ and $\rho(F) = F(y_0)a(w,F)$, where $a(w,F)$ is some function with additional variable w playing the role of a placeholder. For arbitrary y_0, a and w we must have

$$\mathsf{E}_\Psi[(\mathrm{I}\{Y \le y_0\} - \tilde{F}(y_0))a(w,\tilde{F})] = 0.$$

By the substitution property of conditional expectation[15] w here can be replaced by an arbitrary Ψ-measurable random variable W:

$$\mathsf{E}_\Psi[(\mathrm{I}\{Y \le y_0\} - \tilde{F}(y_0))a(W,\tilde{F})] = 0$$

and

$$\mathsf{E}[(\mathrm{I}\{Y \le y_0\} - \tilde{F}(y_0))a(W,\tilde{F})] = 0.$$

Since a here is arbitrary, it follows that

$$\mathsf{E}_{\Psi,\tilde{F}}[\mathrm{I}\{Y \le y_0\} - \tilde{F}(y_0)] = 0,$$

which is equivalent to $\tilde{F} = \mathbb{F}_{\Psi,\tilde{F}}$. □

Theorem 16. *Suppose that in a recursive one-step density forecasting situation each forecast $\tilde{F}_t$, $t = 1,\dots,T$ is Ψ_t-measurable, where $\Psi_t = \sigma(Y_1,\dots,Y_{t-1})$, and that $(P_1,\dots,P_T) \sim U[0,1]^T$, where $P_t = \tilde{F}_t(Y_t)$. Then each forecast $\tilde{F}_t$ is ideally calibrated with respect to Ψ_t.*

Proof. From $(P_1,\dots,P_T) \sim U[0,1]^T$ it follows that $P_t|P_1,\dots,P_{t-1} \sim U[0,1]$. In a recursive one-step density forecasting situation if each $\tilde{F}_t$ is Ψ_t-measurable we have

$$\sigma(P_1,\dots,P_{t-1}) = \sigma(Y_1,\dots,Y_{t-1}) = \Psi_t$$

and thus $P_t|\Psi_t \sim U[0,1]$. By Theorem 14 this means that each $\tilde{F}_t$ is ideally calibrated with respect to Ψ_t. □

Theorem 17. *For any proper scoring rule the forecast, which is ideally calibrated with respect to Ψ, attains the highest expected score among the Ψ-measurable forecasts.*

Proof. Suppose that $\mathbb{F}_\Psi$ is the ideally calibrated forecast and $\tilde{F}$ is some Ψ-measurable forecast. Then by Theorem 6 $\mathsf{E}S(\tilde{F},Y) = \mathsf{E}S(\tilde{F},\mathbb{F}_\Psi)$ and $\mathsf{E}S(\mathbb{F}_\Psi,Y) = \mathsf{E}S(\mathbb{F}_\Psi,\mathbb{F}_\Psi)$. Since S is proper, we have $\mathsf{E}S(\mathbb{F}_\Psi,\mathbb{F}_\Psi) \ge \mathsf{E}S(\tilde{F},\mathbb{F}_\Psi)$ and hence $\mathsf{E}S(\mathbb{F}_\Psi,Y) \ge \mathsf{E}S(\tilde{F},Y)$. □

[15] That is, Ψ-measurable variables can be treated as fixed inside E_Ψ. Here we have to use a generalized form of Theorem 6 above, which can be found in Kallenberg (2002).

References

Amisano, G., and R. Giacomini (2007): "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business and Economic Statistics*, 25(2), 177–190.

Bao, Y., T.-H. Lee, and B. Y. Saltoğlu (2007): "Comparing Density Forecast Models," *Journal of Forecasting*, 26, 203–225.

Berkowitz, J. (2001): "Testing Density Forecasts, With Applications to Risk Management," *Journal of Business & Economic Statistics*, 19(4), 465–474.

Bierens, H. J. (2004): *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge University Press.

Boero, G., J. Smith, and K. F. Wallis (2011): "Scoring Rules and Survey Density Forecasts," *International Journal of Forecasting*, 27(2), 379–393.

Bollerslev, T. (1987): "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return," *Review of Economics and Statistics*, 69(3), 542–547.

Bröcker, J. (2009): "Reliability, Sufficiency, and the Decomposition of Proper Scores," *Quarterly Journal of the Royal Meteorological Society*, 135(643), 1512–1519.

Bröcker, J., and L. A. Smith (2007): "Scoring Probabilistic Forecasts: The Importance of Being Proper," *Weather and Forecasting*, 22, 382–388.

Brockwell, A. E. (2007): "Universal Residuals: A Multivariate Transformation," *Statistics & Probability Letters*, 77, 1473–1478.

Chen, Y.-T. (2011): "Moment Tests for Density Forecast Evaluation in the Presence of Parameter Estimation Uncertainty," *Journal of Forecasting*, 30(4), 409–450.

Chong, Y. Y., and D. F. Hendry (1986): "Econometric Evaluation of Linear Macro-Economic Models," *Review of Economic Studies*, 53(4), 671–690.

Christoffersen, P. F. (1998): "Evaluating Interval Forecasts," *International Economic Review*, 39(4), 841–862.

Clements, M. P. (2006): "Evaluating the Survey of Professional Forecasters Probability Distributions of Expected Inflation Based on Derived Event Probability Forecasts," *Empirical Economics*, 31(1), 49–64.

Clements, M. P., and D. I. Harvey (2010): "Forecast Encompassing Tests and Probability Forecasts," *Journal of Applied Econometrics*, 25(6), 1028–1062.

Clements, M. P., and N. Taylor (2003): "Evaluating Interval Forecasts of High-Frequency Financial Data," *Journal of Applied Econometrics*, 18(4), 445–456.

Corradi, V., and N. R. Swanson (2005): "A Test for Comparing Multiple Misspecified Conditional Interval Models," *Econometric Theory*, 21(05), 991–1016.

——— (2006a): "Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification," *Journal of Econometrics*, 133(2), 779–806.

——— (2006b): "Predictive Density and Conditional Confidence Interval Accuracy Tests," *Journal of Econometrics*, 135(1-2), 187–228.

——— (2006c): "Predictive Density Evaluation," in *Handbook of Economic Forecasting*, ed. by C. W. J. Granger, G. Elliott, and A. Timmermann, vol. 1, chap. 5, pp. 197–286. North-Holland, Amsterdam.

Cox, D. R. (1961): "Tests of Separate Families of Hypotheses," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 105–123, Berkeley. University of California Press.

Cox, D. R. (1962): "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2), 406–424.

Dawid, A. P. (1984): "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society. Series A (General)*, 147(2), 278–292.

DeGroot, M. H. (1962): "Uncertainty, Information, and Sequential Experiments," *The Annals of Mathematical Statistics*, 33(2), 404–419.

DeGroot, M. H., and S. E. Fienberg (1983): "The Comparison and Evaluation of Forecasters," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2), 12–22.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39(4), 863–883.

Diebold, F. X., J. Hahn, and A. S. Tay (1999): "Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns on Foreign Exchange," *Review of Economics and Statistics*, 81(4), 661–673.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–263.

Diebold, F. X., and G. D. Rudebusch (1989): "Scoring the Leading Indicators," *The Journal of Business*, 62(3), 369–391.

Diebold, F. X., A. S. Tay, and K. F. Wallis (1999): "Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters," in *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*, ed. by R. F. Engle, and H. White, pp. 76–90. Oxford University Press, Oxford.

Diks, C., V. Panchenko, and D. van Dijk (2011): "Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails," *Journal of Econometrics*, 163(2), 215–230.

Engelberg, J., C. F. Manski, and J. Williams (2009): "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters," *Journal of Business and Economic Statistics*, 27(1), 30–41.

Engle, R. F., and S. Manganelli (2004): "CAViaR," *Journal of Business & Economic Statistics*, 22(4), 367–381.

Ferguson, T. S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

Galbraith, J. W., and S. van Norden (2011): "Kernel-Based Calibration Diagnostics for Recession and Inflation Probability Forecasts," *International Journal of Forecasting*, 27(4), 1041–1057.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74(6), 1545–1578.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007): "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society: Series B*, 69, 243–268.

Gneiting, T., and A. E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., and R. Ranjan (2013): "Combining Predictive Distributions," *Electronic Journal of Statistics*, 7, 1747–1782.

Granger, C. W. J. (1999): "Outline of Forecast Theory Using Generalized Cost Functions," *Spanish Economic Review*, 1, 161–173.

Granger, C. W. J., and M. H. Pesaran (2000): "A Decision-Theoretic Approach to Forecast Evaluation," in *Statistics and Finance: An Interface*, ed. by W.-S. Chan, W. K. Li, and H. Tong. Imperial College Press.

Hall, S. G., and J. Mitchell (2007): "Combining Density Forecasts," *International Journal of Forecasting*, 23, 1–13.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

Holzmann, H., and M. Eulert (2011): "The Role of the Information Set for Forecasting — With Applications to Risk Management," (unpublished).

Kallenberg, O. (2002): *Foundations of Modern Probability*. Springer, 2 edn.

Kupiec, P. H. (1995): "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, 3(2), 73–84.

Lopez, J. A. (1998): "Methods for Evaluating Value-at-Risk Estimates," *Economic Policy Review*, (October), 119–124.

Mincer, J. A., and V. Zarnowitz (1969): "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, ed. by J. A. Mincer, pp. 3–46. National Bureau of Economic Research.

Mitchell, J., and K. F. Wallis (2011): "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness," *Journal of Applied Econometrics,*, 26(6), 1023–1040.

Murphy, A. H., and R. L. Winkler (1987): "A General Framework for Forecast Verification," *Monthly Weather Review*, 115, 1330–1338.

Pesaran, M. H., and S. Skouras (2002): "Decision-Based Methods for Forecast Evaluation," in *A Companion to Economic Forecasting*, ed. by M. P. Clements, and D. F. Hendry, chap. 11, pp. 241–267. Blackwell.

RiskMetrics (1996): "RiskMetrics(TM) — Technical Document (4th ed.)," Discussion paper, J. P. Morgan/Reuters'.

Sanders, F. (1963): "On Subjective Probability Forecasting," *Journal of Applied Meteorology*, 2, 191–201.

Sarno, L., and G. Valente (2004): "Comparing the Accuracy of Density Forecasts from Competing Models," *Journal of Forecasting*, 23, 541–557.

Shiller, R. J. (1978): "Rational Expectations and the Dynamic Structure of Macroeconomic Models: A Critical Review," *Journal of Monetary Economics*, 4(1), 1–44.

Tsyplakov, A. (2011): "Evaluating Density Forecasts: A Comment," MPRA Paper 32728, University Library of Munich, Germany.

Wallis, K. F. (2003): "Chi-Squared Tests of Interval and Density Forecasts, and the Bank of England's Fan Charts," *International Journal of Forecasting*, 19(2), 165–175.

West, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–1087.

White, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68(5), 1097–1126.